Investigating Selective Prediction Approaches Across Several Tasks in IID, OOD, and Adversarial Settings

Neeraj Varshney, Swaroop Mishra, Chitta Baral
Arizona State University, USA

Selective Prediction

Selective Prediction allows a system to abstain from answering.
A system can typically abstain when its prediction is likely to be incorrect.
This improves the system's **reliability**.

A Selective prediction system comprises of:

- A **predictor** function (f) that gives the model's prediction
- A **selector** function (g) that determines if the system should output the prediction made by the predictor.

Usually, 'g' comprises of a confidence estimator '$\tilde{g}$' that indicates *f's* prediction confidence and a threshold '*th*' that controls the abstention level:

$$(f,g)(x) = \begin{cases} f(x), & \text{if } g(x) = 1 \\ Abstain, & \text{if } g(x) = 0 \end{cases} \qquad g(x) = \mathbb{1}[\tilde{g}(x)) > th]$$

A selective prediction system makes trade-offs between **coverage** and **risk**. For a dataset D, **coverage** at a threshold th is defined as the fraction of total instances answered by the system (where $\tilde{g} > th$) and **risk** is the error on the answered instances:

$$coverage_{th} = \frac{\sum_{x_i \in D} \mathbb{1}[\tilde{g}(x_i)) > th]}{|D|}$$

$$risk_{th} = \frac{\sum_{x_i \in D} \mathbb{1}[\tilde{g}(x_i)) > th] l_i}{\sum_{x_i \in D} \mathbb{1}[\tilde{g}(x_i)) > th]}$$

- With decrease in threshold, coverage will increase, but the risk will usually also increase.
- The overall selective prediction performance is measured by the area under Risk-Coverage curve (AUC).
- **Lower the AUC, the better the selective prediction system as it represents lower average risk across all thresholds.**

Selective Prediction Approaches

Maximum Softmax Probability (MaxProb):

- Usually, the last layer of models has a softmax activation function that gives the probability distribution P(y) over all possible answer candidates Y.
- MaxProb corresponds to the maximum softmax probability across all answer candidates.

Monte-Carlo Dropout (MCD):

- An input is inferred multiple times using different dropout masks and the outputs are aggregated to get the confidence estimate for selective prediction.

Label Smoothing (LS):

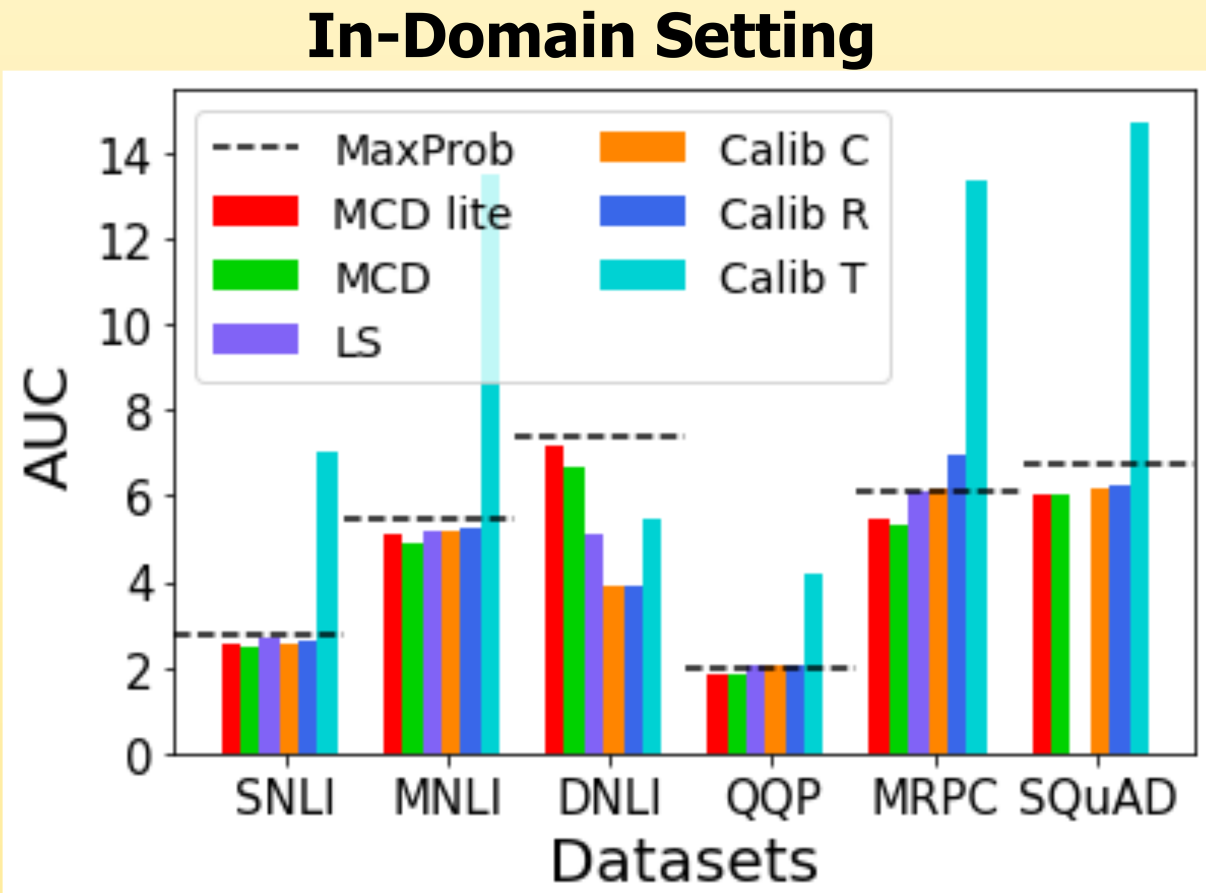
- Cross-entropy loss is calculated with a weighted mixture of target labels instead of one hot 'hard' label during training.
- This prevents the network from becoming overconfident

Calibration:
A held-out dataset is annotated conditioned on the correctness of the model's predictions and a calibrator is trained over this annotated dataset to give the confidence estimate for test instances.

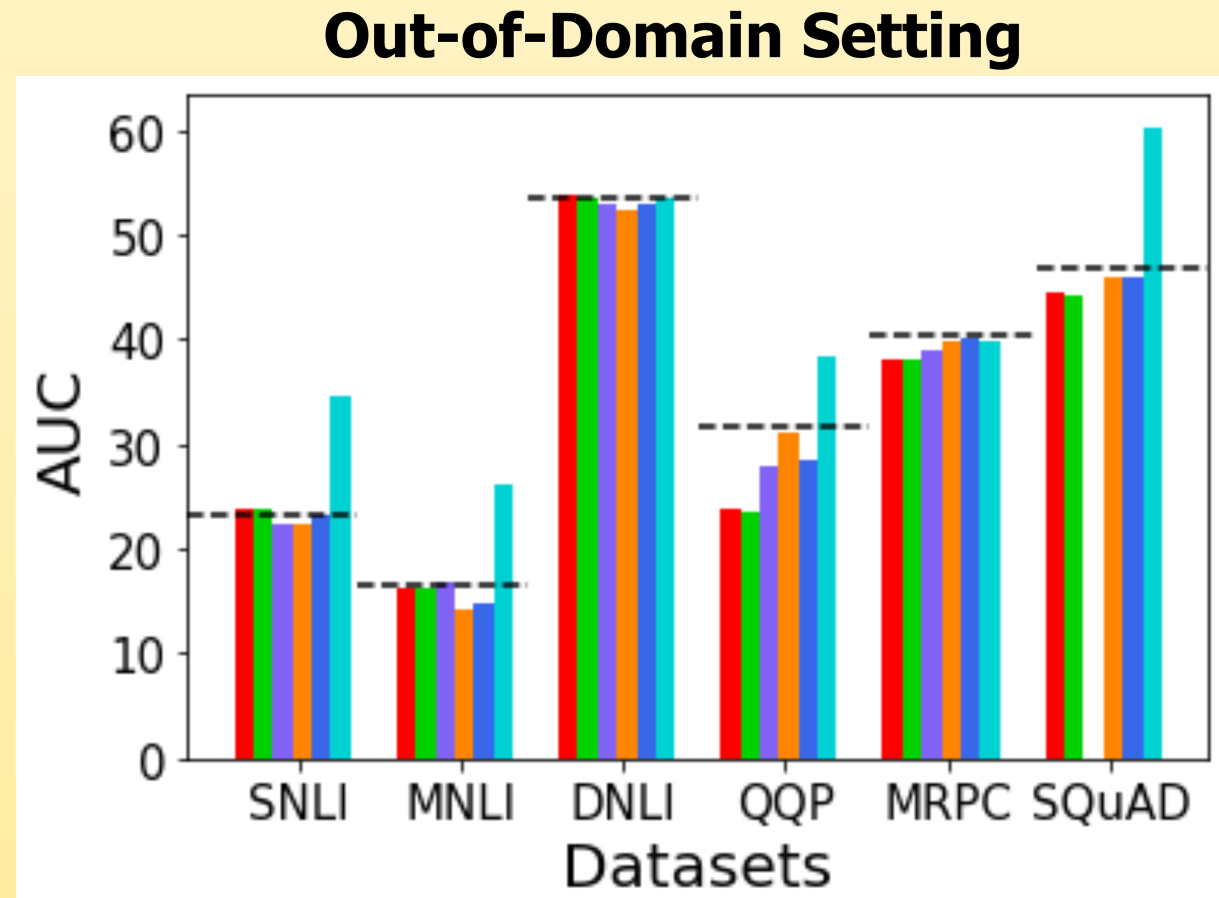
- **Calib C:** Held-out dataset is annotated with two classes (correct as '*positive*' class and incorrect as '*negative*' class), and calibrator is trained on this annotated binary classification dataset. Probability assigned to the positive class by this trained calibrator is used as the confidence for selective prediction.
- **Calib R**: Held-out dataset is annotated on a continuous scale between '0' and '1' instead of categorical labels.
- **Calib T**: A transformer-based model is trained as calibrator that leverages the entire input text for training instead of features derived from it.

Experiments and Results

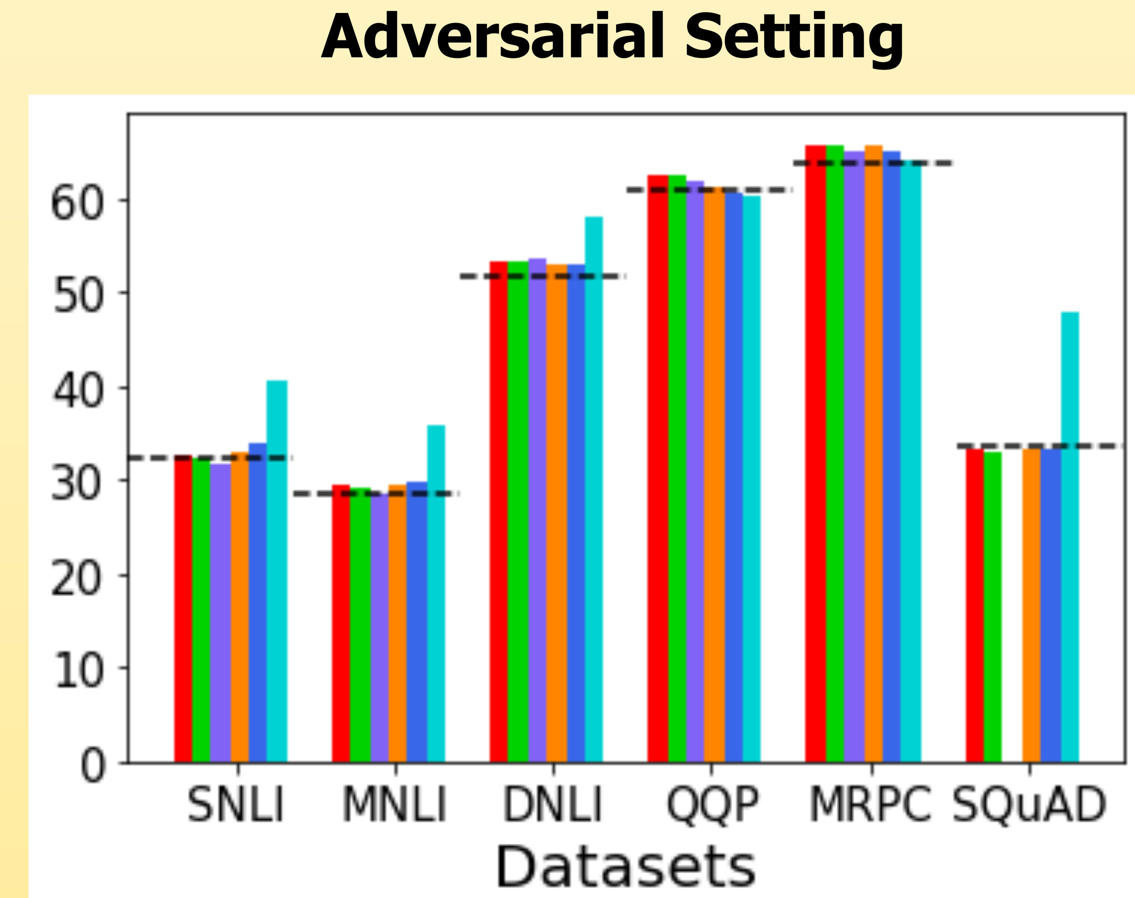
In-Domain Setting



Out-of-Domain Setting



Adversarial Setting



1. **None of the existing selective prediction approaches consistently and considerably outperforms *MaxProb*.**
   - Slight improvement in In-Domain
   - Negligible improvement in Out-of-Domain
   - Performance degradation in Adversarial
2. MCD requires additional computation for multiple inferences and calibration requires additional heldout dataset for training calibrator.
   - In contrast, **MaxProb doesn't require any such additional resources and yet performs well.**
3. **Approaches do not translate well across tasks**
   - *MCD* outperforms all other approaches on Duplicate Detection datasets but does not fare well on the NLI datasets.

Overall, our results highlight that there is a need to develop stronger selective prediction approaches that perform well across multiple tasks (QA, NLI, etc.) and settings (IID, OOD, and ADV) while being resource-efficient.